D74–D78 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 23 October 2016*
doi: 10.1093/nar/gkw945

LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs

Shangwei Ning†, Ming Yue†, Peng Wang†, Yue Liu, Hui Zhi, Yan Zhang, Jizhou Zhang, Yue Gao, Maoni Guo, Dianshuang Zhou, Xin Li and Xia Li*

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Received August 14, 2016; Revised September 27, 2016; Editorial Decision October 08, 2016; Accepted October 19, 2016

ABSTRACT

We describe LincSNP 2.0 (http://bioinfo.hrbmu.edu.cn/LincSNP), an updated database that is used specifically to store and annotate disease-associated single nucleotide polymorphisms (SNPs) in human long non-coding RNAs (lncRNAs) and their transcription factor binding sites (TFBSs). In LincSNP 2.0, we have updated the database with more data and several new features, including (i) expanding disease-associated SNPs in human lncRNAs; (ii) identifying disease-associated SNPs in lncRNA TFBSs; (iii) updating LD-SNPs from the 1000 Genomes Project; and (iv) collecting more experimentally supported SNP-lncRNA-disease associations. Furthermore, we developed three flexible online tools to retrieve and analyze the data. Linc-Mart is a convenient way for users to customize their own data. Linc-Browse is a tool for all data visualization. Linc-Score predicts the associations between lncRNA and disease. In addition, we provided users a newly designed, user-friendly interface to search and download all the data in LincSNP 2.0 and we also provided an interface to submit novel data into the database. LincSNP 2.0 is a continually updated database and will serve as an important resource for investigating the functions and mechanisms of lncRNAs in human diseases.

INTRODUCTION

An abundant class of non-coding RNAs known as long non-coding RNAs (lncRNAs), defined by having a length exceeding 200 nucleotides, have gained widespread attention in recent years (1). LncRNAs are widely encoded by the human genome and perform important functions in a spectrum of biological processes such as genome regulation, cell differentiation and development (2–4). Accumulating evidence indicates that lncRNAs are closely associated with many human diseases (5,6).

In the emerging field of lncRNA research, many researchers have continued to focus on the influence of genetic variants on lncRNA function. A number of single nucleotide polymorphisms (SNPs), the most common type of genetic variant, have been identified in human lncRNA regions and have been shown to be associated with various diseases including cancers (7,8). In order to facilitate the study of lncRNA-related genetic variants, we reported the first version of the LincSNP database (LincSNP 1.0) that allows users to search all known disease-associated SNPs in human lncRNAs, together with their comprehensive functional annotations (9). Although LincSNP 1.0 has provided some useful information for researchers, this database could provide more resources and be more user friendly. For example, LincSNP 1.0 only focused on large intergenic non-coding RNA (lincRNA), a subclass of lncRNAs and did not identify disease-associated SNPs in lncRNA regulatory elements such as transcription factor binding sites (TFBSs). Previous studies have demonstrated that the SNPs in lncRNA TFBSs could affect lncRNA expression, thereby potentially affecting disease susceptibility (10). With the increasing amount of lncRNA and SNP data, there is a great need to update LincSNP 1.0 with more resources and improved tools.

To date, many databases have been built to curate lncRNA-related information, such as NONCODE (11), DIANA-LncBase (12), LNCipedia (13), lncRNAdb (14), LncRNAWiki (15), ChIPBase (16), starBase (17), LncRNADisease (18) and Lnc2Cancer (19). These databases have provided valuable resources for lncRNA-related studies. However, there are very few databases that pay special attention to the relationship between SNPs and human lncRNAs. To our knowledge, only the lncRNASNP database stores lncRNA-related SNPs (20). However, this database focuses mainly on exploring the impact of SNPs on lncRNA structure and function, and only a small number of disease-associated SNPs have been identified in hu-

*To whom correspondence should be addressed. Tel: +86 451 86615922; Fax: +86 451 86615922; Email: lixia@hrbmu.edu.cn
†These authors contributed equally to the work as the first authors.

man lncRNAs. Until now, no specialized resource has been devoted to collecting, storing and distributing disease-associated SNPs in human lncRNAs.

To meet these needs, we have updated LincSNP (9) to version 2.0 (LincSNP 2.0) (Figure 1 and Table 1). In LincSNP 2.0, the numbers of disease-associated SNPs and human lncRNAs have been increased to 809 451 and 244 545, respectively, and the number of types of lncRNA has been increased to 9. For the first time, disease-associated SNPs in lncRNA TFBSs were identified and included in LincSNP 2.0. Furthermore, the number of experimentally supported SNP-lncRNA-disease associations has grown from 3 to 58. In addition to the expansion of the core data sets, both the data search and download functions were improved. In particular, three web-based tools have been developed to facilitate data analysis, extraction and visualization. We hope that researchers will benefit from the greater resources in the updated version of LincSNP 2.0.

IMPROVED CONTENT AND NEW FEATURES

Expanded entries on disease-associated SNPs in human lncRNAs

Recent advances in high-throughput sequencing technology such as RNA-Seq have produced large numbers of lncRNAs (21). There has also been a rapid increase of GWAS data in public databases (22). This information provides us with a great opportunity to identify more disease-associated SNPs in human lncRNAs (Table 1). In LincSNP 2.0, the lncRNA sources have expanded from 1 to 5 databases, including Ensembl (Version 75), LncRBase (Version 1.0), NONCODE (Version 4), LNCipedia (Version 3.1) and GENCODE (Version 19). To provide a universal lncRNA annotation for users, lncRNA transcripts downloaded from different sources were considered to be the same transcript if they had the same positions. Then, each lncRNA transcript was named using serial numbers after the 'LSLNC' symbol. In total, we obtained 244 545 human lncRNAs and their annotations, and the number of types of lncRNA increased to 9 (including lincRNA, 3′ overlapping ncRNA, antisense, processed transcript, exonic, retained intron, sense no exonic, sense intronic and sense overlapping).

The set of human GWAS databases storing disease (traits)-associated SNPs has been expanded from 6 to 8 sources, including dbGaP (23), GAD (24), GWAS Central (25), Johnson and O'Donnell (26), the NHGRI GWAS Catalog (27), PharmGKb (28), GWASdb (Version 2) (22) and GRASP (Version 2) (29). As the integrated strategy in LincSNP 1.0, disease-associated SNPs were selected from original publications with moderate thresholds (P-values $< 1.0 \times 10^{-3}$) and only the most significant SNP was selected in cases where the same SNP could be obtained from different publications (9). In total, 809 451 unique disease-associated SNPs were collected. We also extracted SNPs that had linkage disequilibrium (LD-SNP, $r^2 \geq 0.8$) relationships with disease-associated SNPs from the 1000 Genomes Project (Phase I version 3). After LD analysis by VCFtools (30), ∼11.6 million LD-SNPs were collected in LincSNP 2.0. Finally, we identified 371 647 disease-associated SNPs located in 145 642 human lncRNAs and we identified 1 266 485 LD-SNPs in 168 915 human lncRNAs.

Newly added data on disease-associated SNPs in lncRNA TFBSs

We recently developed a database named SNP@lincTFBS to identify SNPs in potential TFBSs of human lincRNAs (31). The updated LincSNP 2.0 has integrated SNP@lincTFBS as an important resource for the functional annotation of SNPs in lncRNA TFBSs. Briefly, we downloaded ChIP-Seq data sets for human transcription factors and identified the peaks located in the promoter regions of human lncRNAs (5 kb upstream to 1 kb downstream region of the transcription start site for each lncRNA) (32). In total, we identified 5 284 709 TFBSs in the defined promoter regions of 211 928 human lncRNAs. We identified 43 672 disease-associated SNPs in 593 492 TFBSs of 86 495 lncRNAs and we identified 1 250 571 LD-SNPs in 168 915 TFBSs of 123 566 lncRNAs.

Updated entries on experimentally supported SNP-lncRNA-disease associations

To provide a reliable source for the associations between lncRNA-related SNPs and disease, we developed a new page, Linc-Confirm, to store all experimentally supported SNP-lncRNA-disease associations. All experimentally supported SNP-lncRNA-disease associations were manually collected through several steps, as previously described (33–35). First, we downloaded all published literature through searching the PubMed database (36) with a list of keywords (before July 2016), such as 'lncRNA SNP disease,' 'long non-coding RNA SNP disease,' 'lncRNA SNP cancer,' 'lncRNA SNP tumor' and 'long noncoding RNA polymorphism disease.' Second, experimentally supported SNP-lncRNA-disease associations were manually curated from published papers by at least two researchers. We retrieved the lncRNA, SNP and disease name, experimental samples and methods, PubMed ID, paper title and a brief description from the original studies. Third, all selected studies were rechecked for the lncRNA, SNP and disease names and some names were replaced with official or recommended names. In LincSNP 2.0, the number of experimentally supported SNP-lncRNA-disease associations has increased significantly, from 3 to 58 entries.

Linc-Mart tool for data discovery and access

Because of the large increase in the number of data entries, a new data access tool called Linc-Mart was developed to implement a customized data access pipeline for users. There are three options on the Linc-Mart page: selected project (Disease SNP – LncRNA or Disease SNP – LncRNA TFBS), chromosome and lncRNA annotation information. Users can upload an e-mail address, and Linc-Mart will process the file using a series of tunable criterion and filter steps based on the above options.

Linc-Browse tool for customized data views

Compared with LincSNP 1.0, we improved the LincSNP 2.0 architecture by adding the Linc-Browse tool to display important annotation tracks. Linc-Browse is a web-based genome browser that dynamically displays different tracks

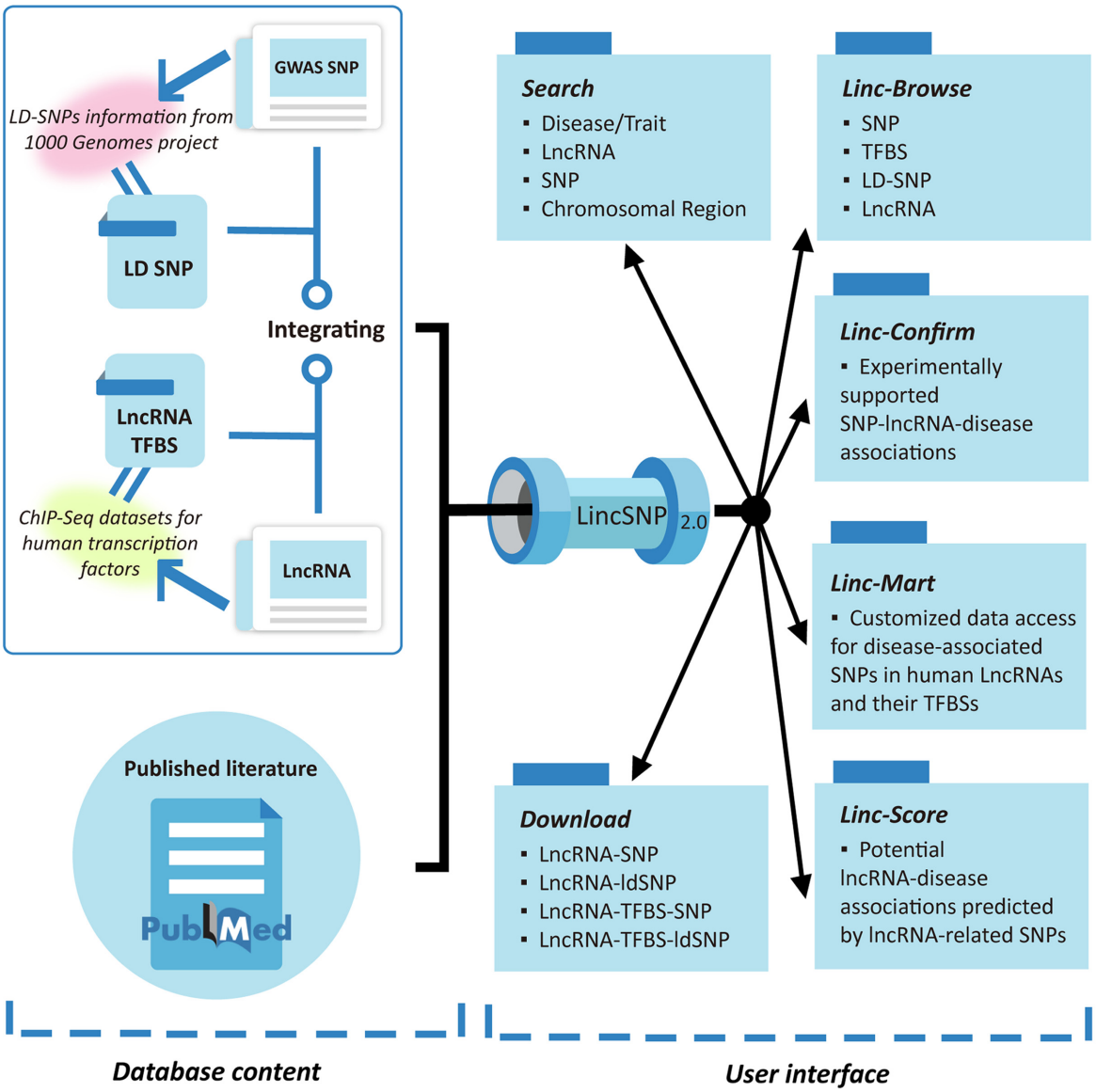

Figure 1. Architecture of LincSNP 2.0.

based on the queried lncRNAs. Linc-Browse provides five tracks (Reference sequence, lncRNA, SNP, LD-SNP and TFBS) that show the elements around the lncRNAs.

Linc-Score page for predicting disease-associated lncRNAs

In LincSNP 2.0, we developed a page called Linc-Score, which was used to predict potential lncRNA-disease associations based on genetic mutations. For each lncRNA and each specific disease, we counted the number of disease-associated SNPs located in this lncRNA region and TFBSs, including both disease-associated SNPs and their LD SNPs. We then calculated the top three potential lncRNA-disease associations for each lncRNA based on the number of disease-associated SNPs. We hope that this direct calculation can capture the lncRNAs most likely to be involved in specific diseases, providing disease lncRNA candidates for researchers.

DATABASE CONSTRUCTION AND IMPROVED USER INTERFACE

All data in LincSNP 2.0 were stored and managed using MySQL (version 5.5.58). The web interfaces were upgraded by applying Linux, Apache, MySQL and PHP (pre hypertext processor) (LAMP) technologies. The LincSNP 2.0 database is freely available at http://bioinfo.hrbmu.edu.cn/LincSNP or http://210.46.80.146/lincsnp. In addition, for the convenience of users who have used LincSNP 1.0, the old version is still in service. Researchers can enter it by clicking on the gateway in the LincSNP 2.0 homepage or go directly to http://210.46.85.180:8080/LincSNP.

LincSNP 2.0 provides a user-friendly web interface that enables users to search, browse and download data in a few easy steps. From the 'Search' page, users can search by Disease or Trait name, LncRNA transcript name and alias, SNP rs number or Chromosomal region. Three flexible on-

Table 1. Content and entries of LincSNP 2.0

Database content	LincSNP 1.0	LincSNP 2.0	Fold increase
SNP	128 407 SNPs	809 451 SNPs	6.3
dbGaP	Yes	Yes	
GAD	Yes	Yes	
GWAS Central	Yes	Yes	
Johnson and O'Donnell	Yes	Yes	
NHGRI GWAS Catalog	Yes	Yes	
PharmGKb	Yes	Yes	
GWASdb	No	Yes	
GRASP	No	Yes	
lncRNA	5804 lincRNAs	244 545 lncRNAs	42.1
Ensembl	Yes	Yes	
NONCODE	No	Yes	
LNCipedia	No	Yes	
LncRBase	No	Yes	
GENCODE	No	Yes	
LD-SNP	1.5 million	11.6 million	7.7
Origin	HapMap	1000 Genomes Project	
TFBS	No	162 human TFs	
Origin	No	ChIP-Seq data	
Validated data	3 entries	58 entries	19.3
Origin	PubMed	PubMed	

line tools, Linc-Mart, Linc-Browse and Linc-Score, were established to retrieve and analyze the data in LincSNP 2.0. LincSNP 2.0 is totally open source, and users can obtain all data from the 'Download' page. LincSNP 2.0 also offers a submission page that enables researchers to submit novel experimentally supported SNP-lncRNA-disease associations, and a detailed tutorial showing how to use LincSNP 2.0 is available on the 'Help' page.

CONCLUSIONS AND FUTURE DEVELOPMENT

When we developed the first version of the LincSNP database (LincSNP 1.0), only a limited number of disease-associated SNPs had been identified in human lncRNAs. With the very fast growth of identified lncRNAs and disease-associated SNPs, there is a great need to update the LincSNP database. In LincSNP 2.0, more disease-associated SNPs in human lncRNAs were identified and annotated. To improve the functions of data processing and database access, three web-based tools, Linc-Mart, Linc-Browse and Linc-Score, were developed. Moreover, we used ChIP-Seq data sets to identify disease-associated SNPs in the TFBSs of lncRNAs. We expect that the number of disease-associated SNPs mapped to lncRNAs and their TFBSs will continue to increase rapidly in the future releases of the LincSNP database. We will continually maintain and update the LincSNP database and integrate more data sets into the LincSNP database, such as cancer genomics data and clinical information, which will improve our understanding of the function of lncRNAs in human diseases.

FUNDING

National High Technology Research and Development Program of China [863 Program, 2014AA021102]; National Program on Key Basic Research Project [973 Program, 2014CB910504]; National Natural Science Foundation of China [91439117, 61473106 and 31401090]; Creative Research Groups of the National Natural Science Foundation of China [81421063]; Postdoctoral Science Foundation of China [2015M571432, 2016T90308 and LBH-Z14148]; Harbin Special Funds for Innovative Talents of Science and Technology Research Project [RC2016QN003028]; Yu Weihan Outstanding Youth Training Fund of Harbin Medical University, and Key Laboratory of Cardiovascular Medicine Research (Harbin Medical University), Ministry of Education. Funding for open access charge: National High Technology Research and Development Program of China [863 Program, 2014AA021102]; National Program on Key Basic Research Project [973 Program, 2014CB910504]; National Natural Science Foundation of China [91439117, 61473106 and 31401090]; Creative Research Groups of the National Natural Science Foundation of China [81421063]; Postdoctoral Science Foundation of China [2015M571432, 2016T90308 and LBH-Z14148]; Harbin Special Funds for Innovative Talents of Science and Technology Research Project [RC2016QN003028]; Yu Weihan Outstanding Youth Training Fund of Harbin Medical University, and Key Laboratory of Cardiovascular Medicine Research (Harbin Medical University), Ministry of Education.

Conflict of interest statement. None declared.

REFERENCES

1. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
2. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
3. Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, **14**, 699–712.
4. Fatica,A. and Bozzoni,I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.*, **15**, 7–21.
5. Huarte,M. (2015) The emerging role of lncRNAs in cancer. *Nat. Med.*, **21**, 1253–1261.
6. Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.
7. Jin,G., Sun,J., Isaacs,S.D., Wiley,K.E., Kim,S.T., Chu,L.W., Zhang,Z., Zhao,H., Zheng,S.L., Isaacs,W.B. *et al.* (2011) Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis*, **32**, 1655–1659.

8. Li,D., Song,L., Wen,Z., Li,X., Jie,J., Wang,Y. and Peng,L. (2016) Strong evidence for LncRNA ZNRD1-AS1, and its functional Cis-eQTL locus contributing more to the susceptibility of lung cancer. *Oncotarget*, **7**, 35813–35817.
9. Ning,S., Zhao,Z., Ye,J., Wang,P., Zhi,H., Li,R., Wang,T. and Li,X. (2014) LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. *BMC Bioinformatics*, **15**, 152.
10. Jendrzejewski,J., He,H., Radomska,H.S., Li,W., Tomsic,J., Liyanarachchi,S., Davuluri,R.V., Nagy,R. and de la Chapelle,A. (2012) The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 8646–8651.
11. Xie,C., Yuan,J., Li,H., Li,M., Zhao,G., Bu,D., Zhu,W., Wu,W., Chen,R. and Zhao,Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98–D103.
12. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Reczko,M., Maragkakis,M., Dalamagas,T.M. and Hatzigeorgiou,A.G. (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.*, **41**, D239–D245.
13. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, D174–D180.
14. Quek,X.C., Thomson,D.W., Maag,J.L., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
15. Ma,L., Li,A., Zou,D., Xu,X., Xia,L., Yu,J., Bajic,V.B. and Zhang,Z. (2015) LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res.*, **43**, D187–D192.
16. Yang,J.H., Li,J.H., Jiang,S., Zhou,H. and Qu,L.H. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res.*, **41**, D177–D187.
17. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.
18. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
19. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L. *et al.* (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
20. Gong,J., Liu,W., Zhang,J., Miao,X. and Guo,A.Y. (2015) lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.*, **43**, D181–D186.
21. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
22. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
23. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
24. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
25. Beck,T., Hastings,R.K., Gollapudi,S., Free,R.C. and Brookes,A.J. (2014) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.*, **22**, 949–952.
26. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
27. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
28. Altman,R.B. (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.*, **39**, 426.
29. Eicher,J.D., Landowski,C., Stackhouse,B., Sloan,A., Chen,W., Jensen,N., Lien,J.P., Leslie,R. and Johnson,A.D. (2015) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, **43**, D799–D804.
30. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
31. Ning,S., Zhao,Z., Ye,J., Wang,P., Zhi,H., Li,R., Wang,T., Wang,J., Wang,L. and Li,X. (2014) SNP@lincTFBS: an integrated database of polymorphisms in human LincRNA transcription factor binding sites. *PLoS One*, **9**, e103851.
32. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
33. Hsu,S.D., Tseng,Y.T., Shrestha,S., Lin,Y.L., Khaleel,A., Chou,C.H., Chu,C.F., Huang,H.Y., Lin,C.M., Ho,S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.
34. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
35. Li,Y., Qiu,C., Tu,J., Geng,B., Yang,J., Jiang,T. and Cui,Q. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.
36. Coordinators,N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.